

Cartographic Representation of Spatial Data Quality Parameters in Volunteered Geographic Information

Farid Karimipour*, Roya Esmaeili**, Gerhard Navratil***

* Department of Surveying and Geomatics Engineering, College of Engineering, University of Tehran, Iran

** Department of Geomatics Engineering, Graduate University of Advanced Technology, Kerman, Iran

*** Department of Geoinformation and Cartography, Technical University of Vienna, Austria

Abstract. Volunteered geographic information (VGI) are constantly being added, edited or removed by the users, so their quality is not static. As VGI users do not necessarily have high spatial knowledge, system administrators control the quality of information in order to provide the users with the appropriate datasets. However, quality of spatial data has several parameters, so the appropriate data may differ from one application to another. Unlike the geographic communities, presenting the standard metadata statements is not so efficient, as they may not be familiar for VGI users. In this paper, we propose providing VGI users with the spatial data quality parameters through simple cartographic representations and let them decide on appropriateness of the datasets for the application at hand. The users select the desired quality parameters as well as the visualization element (e.g. color, line thickness, intensity, style, etc.) to classify the datasets. The datasets are represented by the selected element based on their metadata information, which help the users to visually evaluate the quality of datasets.

Keywords: Volunteered geographic information (VGI), spatial data quality, visual representation

1. Introduction

Recent advances in spatial data collection technologies and online services have dramatically increased the contribution of ordinary people to produce,

share and use geographic information. A growing number of cell phones, digital cameras, PDAs and other hand-held devices are equipped with georeferenced data collection technologies, made it possible for ordinary people to collect spatial data, which are then shared and disseminated on the internet using the web map services. It has led to a huge source of spatial data termed as Volunteered Geographic Information (VGI) by Mike Goodchild (Goodchild 2007). There are several VGI environments, such as OpenStreetMap (OSM), Wikimapia, etc. whose data are provided by the users.

Volunteered geographic information are constantly being added, edited or removed by the users. Like other crowd-source data environment (e.g., Wikipedia), existing data can be improved by users. Thus, quality of volunteered geographic information is not static. As VGI users do not necessarily have high spatial knowledge, they cannot decide on quality of available datasets. Therefore, system administrators control the quality of information in order to provide the users with the appropriate datasets. However, quality of spatial data has several parameters such as spatial accuracy, attribute accuracy, completeness, logical consistency and updateness. Therefore, the appropriate data may differ from one application to another depending on the quality parameters that are important for the given application. For example, for route planning, a more complete dataset with less spatial accuracy may be more relevant than an incomplete dataset with high spatial accuracy. It is very common in geographic communities to evaluate the relevancy of datasets for an application based on metadata, which expresses different aspects of quality of datasets. In case of VGI, however, the users are not experts and do not necessarily have enough spatial knowledge to communicate with the standard metadata statements.

In this paper, we propose providing the VGI users with the spatial data quality parameters through simple cartographic representations and let them decide on evaluating the appropriateness of datasets for the application at hand. The users select the desired quality parameters as well as the visualization element (e.g. color, line thickness, intensity, style, etc.) to classify the datasets. The datasets are represented by the selected element based on their metadata information, which helps the users to visually evaluate the quality of datasets.

The rest of the paper is structured as follows: Section 2 contains an overview of quality issues in VGI conducted in recent years. Approaches for quality assurance, quality assessment and quality representation in VGI are discussed in this regard. The proposed approach for representing spatial data quality parameters in VGI is described in Section 3 and is implement-

ed for a case study in Section 4. Finally, Section 5 concludes the paper and proposes ideas for future work.

2. Quality Issues in VGI

As the amount, variety and usage of spatial data is increasing, spatial data quality is getting more attention (Ather 2009). "Unlike the geographic information produced by mapping agencies and corporations, VGI carries no guarantees of accuracy" (Goodchild 2009a), so their quality and reliability is questionable.

The risks of using poor quality VGI are primarily the same as the risks of using poor quality data from an official or commercial supplier – the source of the data will not affect the results of using the data. The key difference might be that an official agency or commercial vendor could possibly be held legally accountable for their data, though in practice, this hardly ever happens because of disclaimers of liability (Cooper et al. 2012).

As VGI is mostly based on human experience of geography, deploying perception-based parameters to express their spatial quality is more efficient than measurement-based parameters used in case of official spatial data. Navratil (2009) proposed expressing quality of spatial data with possibility distributions instead of precise numbers. De Longueville *et al.* (2010) introduced the concept of *degree of truth* to describe object models with vagueness. Instead of evaluating an object to have a given characteristic, it is expressed with the degree of truth that an object tends to have this characteristic (De Longueville et al. 2010). Flanagan and Metzger believe that *credibility*, as a perceptual variable, is adequate for evaluating collaborative productions. "Although there is no clear definition of credibility, it is generally thought to be the believability of a source or message, which is composed of two primary dimensions: trustworthiness and expertise" (Flanagan & Metzger 2008).

Exel expressed that according to the users experience and local knowledge, the reputation of users of VGI websites can be assessed (Exel et al. 2010). Wikimapia has ranked its users according to the level of experience, number of edits, and number of objects, pictures, etc. they have uploaded. For example, a new comer is ranked as lower level (level zero), which means he can only add objects, but do not have the right to delete any object on the map. As he gains more experience in mapping, his level raises providing him with more authority.

Generally, data quality management has two major components: (1) *Quality assurance* that controls the quality of data during the data creation; and

(2) *Quality assessment* that evaluates the quality of the produced data, whose result is organized in the form of metadata (Goodchild & Li 2012).

2.1. Quality Assurance in VGI

Goodchild presented three approaches for assuring the quality of VGI (Goodchild & Li 2012):

- **Crowd-sourcing approach:** Information provided by a group of people tends to be more accurate than by a single individual. This approach has been deployed by OSM in an online service called "*potlatch*" where users can modify the existing data. Furthermore, OSM users can mark the detected errors in *OpenStreetBugs* to be modified by other users. This approach may not be so useful for geographic domain because there will be many spatial errors despite of many volunteers in an area.
- **Social approach:** Gate-keepers, who are the administrators or high level users, check new data in order to avoid gross errors, vandalism, etc. In OSM, the Data Working Group (DWG) solves the problems such as conflicts in data provided by different users, as well as probable vandalism and violation. In Wikimapia, the vandals found by the low-level users are introduced to the gate-keepers, who have the rights to limit or even block their activities.
- **Geographic approach:** This approach has been specifically designed for geographic data and could be automated. It suggests checking the geographic data according to some rules. For example, the data close to each other must be consistent. This is also true when correlating datasets, e.g., water flows downhill. The *keepright.at* website automatically detects some of the errors exist in the OSM data and highlight them by different symbols to be corrected by the users.

2.2. Quality Assessment in VGI

Having produced the data, their quality is assessed by using metadata or through comparing it with a reference data.

2.2.1. Metadata

The vagueness of crowd-source data could be determined by two types of metadata (De Longueville et al. 2010):

- **User-encoded vagueness metadata:** The user may contribute in giving more information about the collected data, although it is not fully reliable (Goodchild 2008). For example, the user may specify the spatial resolution of the image from which a dataset has been digitized.
- **System-created vagueness metadata:** The system itself store some parameters related to the quality of data, like the scale in which the data has been added.

Despite the importance of metadata for the VGI users, almost no metadata exists for projects such as OpenStreetMap and Google Earth (Goodchild 2009b).

On the other hand, the existing metadata standards may not be relevant for collaborative volunteered data. In geospatial web, relative quality of datasets that are being integrated is very important. Goodchild introduced a binary user-centric metadata, called *metadata 2.0*. In addition to the single quality of the data, metadata 2.0 describes the ability of two datasets to work together (Goodchild 2008).

2.2.2. Comparison

A common approach to measure the spatial quality is comparing the data with a reference. The main issue here is choosing a proper reference dataset (Haklay 2010), especially in VGI where no reference dataset may be available at all (Ciepluch et al. 2011).

Goodchild and Hunter (1997) proposed a method for evaluating the positional accuracy of a dataset comparing with a reference. A buffer with a certain width is created around the reference objects; and the proportion of the tested dataset lies within the buffers is calculated (*Figure 1*). The level of accuracy of the tested dataset depends on the size of the buffer chosen (Goodchild & Hunter 1997). This method has been used to compare OSM data with OS (Ather 2009) and HMGS (Kounadi 2009).

2.3. Quality Presentation

Having evaluated the quality of data, this information is presented to the users in order to help them to assess the fitness of the data for their use. Data producers often provide metadata describing different aspects of the quality of the datasets. A metadata record is a file of information, which captures the basic characteristics of a data or information resource.

In VGI, several people with different knowledge and expertise contribute in acquiring huge amounts of data. It results in datasets with different charac-

teristics. Furthermore, VGI users do not necessarily have high spatial knowledge to evaluate the quality of datasets. Therefore, system administrators control the quality of information in order to provide the users with the appropriate datasets. However, quality of spatial data has several parameters, so the appropriate data may differ from one application to another. It is a situation where different users prefer different datasets, but they are not expert enough to select it based on technical metadata statements.

Currently, produced metadata has limitations in term of communications media for non-expert users and expert users (De Longueville et al. 2010). Metadata often contains technical descriptions and terminologies interpreted by experts. Therefore, it is not so efficient for non-expert users of VGI (Devillers et al. 2002).

There have been ideas proposed for presenting the quality information to the users of crowd-source environments through familiar concepts. For instance, Wikipedia designed an extension called *WikiTrust* to visually present the quality of its articles to the users. Contents of the articles are analyzed based on their stability, creditability, history, etc., whose result is demonstrated in the color intensity of the background. This idea is adapted in the next section to present the quality information to VGI users.

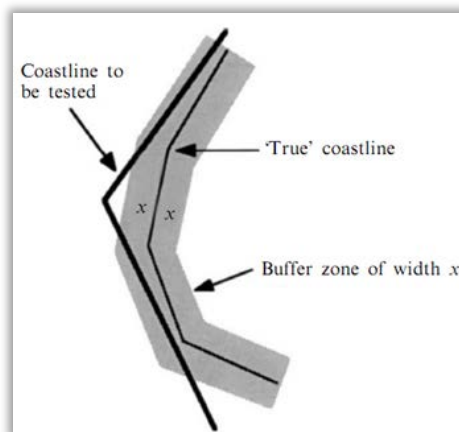


Figure 1. Buffer method to evaluate the positional accuracy of a dataset comparing with a reference (Goodchild & Hunter 1997)

3. Proposed Approach to Represent Spatial Data Quality Information in VGI

In this section, we adapt the idea used in *WikiTrust* to provide the non-expert VGI users with the quality information of spatial data, in order to help them evaluating the datasets for the application at hand. Having determined the quality parameters of the datasets, the features are demonstrated by a certain visual element so that the user has an understanding of their quality. The visual elements can be used as follows (*Figure 2*):

- *Color classification*: The datasets are classified to different quality classes. Then, the datasets of each class are shown in different colors. For example, the datasets with the highest, medium and the lowest quality are drawn in, respectively, green, yellow and red. This can be used for point, line and polygon feature datasets.
- *Color intensity*: The datasets are ordered based on their quality and they are shown by different color intensity. For example, the datasets with the highest and the lowest quality are drawn in, respectively, dark blue and light blue. This can be used for point, line and polygon feature datasets.
- *Feature Size*: The datasets are ordered based on their quality and they are symbolized by different size. For point features, it means different symbol size; and for line features, it is different line thickness. For polygon features, it can be adapted as differentiation in hatching intensity.

4. Implementation

This section describes the results of an implementation developed based on the proposed approach. First, data collection process is introduced. Quality assessment and presentation of the collected data are presented afterwards.

4.1. Data Collection

Ten planimetric maps were produced for a small area, shown in *Figure 3*, using different data collection methods: walking, metering, GPS marking, GPS tracking, digitizing and surveying using total station (*Figures 4 and 5*). In order to have datasets with different spatial qualities (limited here to positional accuracy and completeness), the data collection was performed by different users.

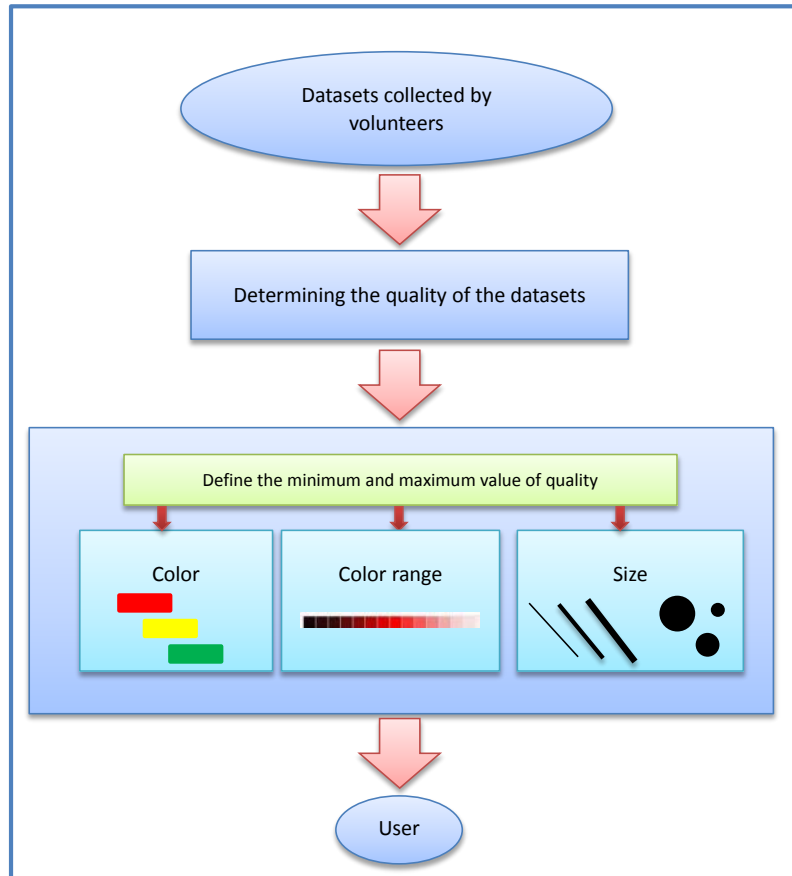


Figure 2. The proposed approach for representation of spatial quality information to the VGI user



Figure 3. The study area

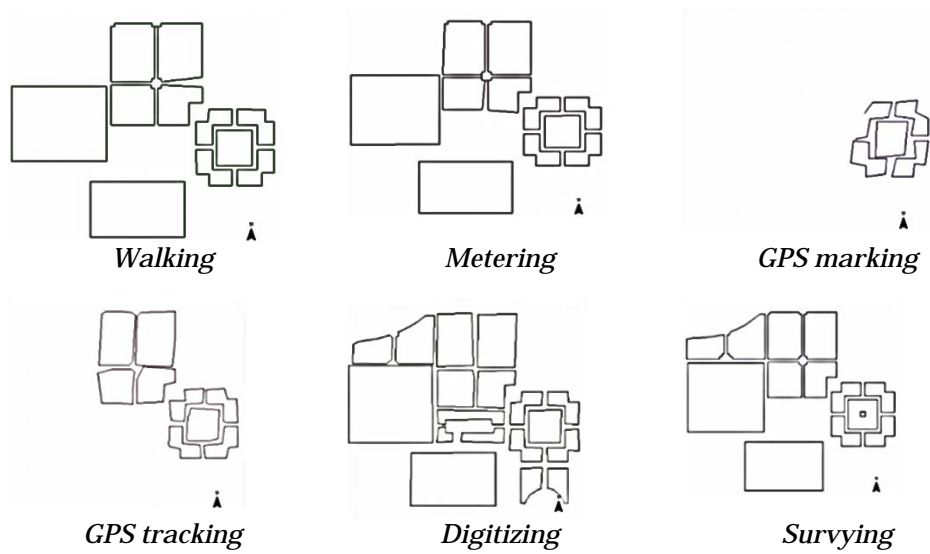


Figure 4. Examples of the produced maps from the study area

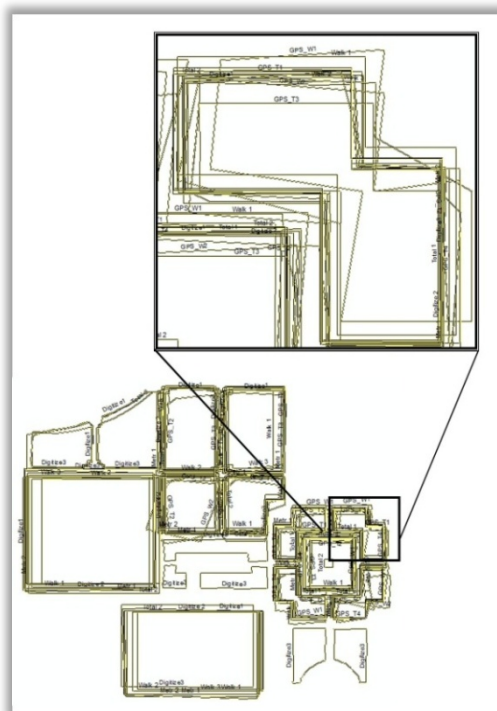


Figure 5. Overlay of the ten datasets collected from the study area

4.2. Quality Assessment

For each datasets, positional accuracy and completeness were assessed as follows:

- **Positional accuracy:** Since there is no reference data to assess the positional accuracy of the datasets, we obtained a relative positional accuracy for each one: First, an initial positional accuracy was assigned to each dataset depending on the data collection method and the instruments used. This initial value is considered as the weight where a weighted average coordinate were computed for each point using its coordinate in all of the datasets in which the point has appeared. For each point in each dataset, we calculated its deviation from the average. Finally, the average of all the deviations calculated for each dataset is assigned to that dataset as its positional accuracy.
- **Completeness:** Again, we calculated a relative completeness parameter for the datasets: The union of all points appeared in all of the datasets was supposed to be the complete data (we assumed no straight line is split into several segments in any of the datasets). Dividing the number of points of each dataset to all points yields its completeness.

4.3. Quality Presentation

An ArcGIS extension was developed to visually represent the quality information (i.e., positional accuracy and completeness) assessed for each dataset to the user. The user selects a number of datasets as well as the desired quality parameter (positional accuracy or completeness). In case of color intensity, a base color is selected by the user; then the desired quality parameter of the selected datasets are distributed over the gray scale of 0 to 255. In case of using symbol size, the minimum and maximum symbol size are set by the user (*Table 1*); the symbol size of the selected datasets are distributed over this range according to the selected quality parameter.

5. Conclusion and Future Work

This paper presents the important issues related to quality management in Volunteered geographic information (VGI). Quality assurance, assessment and representation of VGI were discussed in this regard. Especially, we propose providing the users with the spatial quality information through visual elements. As VGI users do not necessarily have high spatial knowledge, this approach helps them to evaluate and compare the available dataset based on quality parameters important in his current application.

- Ciepluch B, Mooney P, Winstanley A (2011) Building Generic Quality Indicators for OpenStreetMap. 19th Annual GIS Research UK (GISRUK).
- Cooper A, Coetzee S, Kourie D (2012) Volunteered Geographical information – The Challenges. PoPositionIT, pp. 34-38.
- De Longueville B, Ostländer N, Keskitalo C (2010) Addressing Vagueness in Volunteered Geographic Information (VGI) – A Case Study. International Journal of Spatial Data Infrastructures Research, Vol. 5.
- Devillers R, Gervais M, Bédard Y, Jeansoulin R (2002) Spatial Data Quality: From Metadata to Quality Indicators and Contextual End-User Manual. OEEPE/ISPRS Joint Workshop on Spatial Data Quality Management, pp. 21-22.
- Exel M, Dias E, Fruijt S (2010) The Impact of Crowdsourcing on Spatial Data Quality Indicators. GIScience. - Zurich, Switzerland, University of Zurich.
- Flanagin A, Metzger M (2008) The Credibility of Volunteered Geographic Information. GeoJournal, pp. 137-148.
- Goodchild M, Hunter G (1997) A Simple Positional Accuracy Measure for Linear Features. International Journal of Geographical Information Science, Vol. 11, pp. 299-306.
- Goodchild M (2007) Citizens as Sensors: the World of Volunteered Geography. GeoJournal, pp. 211-221.
- Goodchild M (2008) Spatial Accuracy 2.0. The 8th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences.
- Goodchild M (2009a) Geographic Information Systems and Science: Today and Tomorrow. The Association of American Geographers, pp. 3-9.
- Goodchild M (2009b) NeoGeography and the Nature of Geographic Expertise. Journal of Location Based Services, Vol. 3, pp. 82-96.
- Goodchild M, Li L (2012) Assuring the Quality of Volunteered Geographic Information. Spatial Statistics, Vol. 1, pp. 110-120.
- Haklay M (2010) How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. Environment and Planning B: Planning and Design, Vol. 37, pp. 682-703.
- Kounadi O (2009) Assessing The Quality of OpenStreetMap Data. M.Sc. Thesis, University College of London.
- Navratil G. (2009) Modeling Data Quality with Possibility Distributions, In: Stein A., Shi W. & Bijker W. (Eds), Quality Aspects in Spatial Data Mining, CRC Press, Boca Raton, pp. 91-99.